Pairwise alignment incorporating dipeptide covariation Supplemental Appendix: Estimating probabilities from counts with a prior of uncertain reliability

Gavin E. Crooks*, Richard E. Green and Steven E. Brenner
Dept. of Plant and Microbial Biology,
111 Koshland Hall #3102,
University of California, Berkeley,
CA 94720-3102, USA

August 29, 2005

A common problem is that of estimating a discrete probability distribution, $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$, given a limited number of samples drawn from that distribution, summarized by the count vector $n = \{n_1, n_2, \ldots, n_k\}$, and a reasonable *a priori* best guess for the distribution $\theta \approx \pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$. (For a general introduction, see Durbin *et al.* 1998.) This guess may simple be the uniform probability, $\pi_i = 1/k$, which amounts to asserting that, as far as we know, all possible observations are equally likely. At other times, we may know some some more detailed approximation to the distribution θ.

For example, in the present case we wish to estimate the probabilities of substituting a pair of amino acid residues by another residue pair, given the number of times that this substitution has been observed in the training dataset. This probability is hard to estimate reliably since the distribution is very large with $20^4 = 160,000$ dimensions. Moreover, many of the possible observations occur very rarely. However, substitutions at different sites are not strongly correlated, and therefore we may approximate the doublet substitution probabilities by a product of single substitution probabilities. Since the dimensions of these marginals are relatively small we can accurately estimate them from the available data, and thereby construct a reliable and reasonable initial guess for the full doublet substitution distribution.

In the common and conventional pseudocount approach, we assume that the distribution π was estimated from A previous observations. These pseudocounts, $\alpha_i = \pi_i A$, are then proportionally averaged with the real observations ($N = \sum_i n_i$) to provide an estimate of θ;

$$\theta_i = \frac{\alpha_i + n_i}{A + N}. \qquad (1)$$

This prescription is intuitively appealing. When the total number of real counts is much less than the number of pseudocounts ($N \ll A$) the prior dominates, and the estimated distribution is determined by our initial guess, $\theta \approx \pi$. In the alternative limit that the real observations greatly outnumber the pseudocounts ($N \gg A$) the estimated distribution is given by the frequencies $\theta_i = n_i/N$. However, it is not immediately obvious how to select A, although many heuristics have been proposed, including $A = 1$, $A = k$ (Laplace), and $A = \sqrt{N}$ (e.g. Lawrence *et al.*, 1993; Durbin *et al.*, 1998; Nemenman *et al.*, 2001). Essentially, this total pseudocount parameter represents our confidence that the initial guess $\theta \approx \pi$ is accurate, since the larger the total pseudocount the more data is required to overcome this assumption.

Within a Bayesian approach we can avoid this indeterminacy by admitting that, *a priori*, we do not know how confidant we are that π approximates θ. The probability $P(n|\theta)$ of independently sampling a particular set of observations, n, given the underlying sampling probability, θ, follows the multinomial distribution, the multivariate generalization of the binomial distribution;

$$\mathcal{M}(n|\theta) = \frac{1}{M(n)} \prod_{i=1}^{k} \theta_i^{n_i}, \quad M(n) = \frac{\prod_i n_i!}{(\sum_i n_i)!}. \qquad (2)$$

The prior probability of the sampling distribution $P(\theta)$ is typically modeled with a Dirichlet distribution,

$$\mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{k} \theta_i^{(\alpha_i - 1)}, \quad Z(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(A)}. \qquad (3)$$

where $\sum_i \theta = 1$, $\alpha_i > 0$ and $A = \sum_i \alpha_i$. Note that the mean of a Dirichlet is

$$\mathrm{E}[\theta_i] = \frac{\alpha_i}{A}. \qquad (4)$$

Therefore, we may fix the parameters of the Dirichlet prior by equating our initial guess, π, with the mean prior distribution: $\pi = \alpha/A$. If we can fix the scale factor A, then we can combine the prior and observations using Bayes' theorem.

$$P(\theta|n) = \frac{P(n|\theta)P(\theta)}{P(n)}. \qquad (5)$$

Because the multinomial and Dirichlet distributions are naturally conjugate, the posterior distribution $P(\theta|n)$ is also Dirichlet.

$$P(\theta|n) \quad \propto \quad \mathcal{M}(n|\theta)\mathcal{D}(\theta|A\pi)$$

*gec@compbio.berkeley.edu

$$
\begin{aligned}
&\propto \prod_{i=1}^{k} \theta_i^{(A\pi_i+n_i-1)}, \\
&= \mathcal{D}(\theta|A\pi+n) \qquad (6)
\end{aligned}
$$

The last line follows because the product in the previous line is an unnormalized Dirichlet with parameters $(A\pi + n)$, yet the probability $P(\theta|n)$ must be correctly normalized.

Given multinomial sampling and a Dirichlet prior, the probability of the data is given by the under-appreciated multivariant negative hypergeometric distribution (Johnson & Kotz, 1969; Durbin *et al.*, 1998, Eq. 11.23);

$$
\begin{aligned}
P(n) &= \int d\theta\, P(n|\theta)P(\theta), \\
&= \int d\theta\, \mathcal{M}(n|\theta)\mathcal{D}(\theta|A\pi), \\
&= \frac{1}{Z(A\pi)}\frac{1}{M(n)}\int d\theta \prod_{i=1}^{20}\theta_i^{(A\pi_i+n_i-1)}, \\
&= \frac{Z(A\pi+n)}{Z(A\pi)M(n)} \equiv \mathcal{H}'(n|A\pi+n). \qquad (7)
\end{aligned}
$$

Again, the last line follows because the product in the previous line is an unnormalized Dirichlet with parameters $(A\pi + n)$. Therefore, the integral over θ must be equal to the corresponding Dirichlet normalization constant, $Z(A\pi + n)$. Note that, confusingly, the negative hypergeometric distribution is sometimes called the inverse hypergeometric, an entirely different distribution, and vice versa.

Since we do know a reasonable value for the scale factor A we cannot use a simple Dirichlet prior. As an alternative, we explicitly acknowledge our uncertainly about A by building this indeterminacy into the prior itself. Rather than a single Dirichlet, we use the Dirichlet mixture;

$$
P(\theta|\pi) = \int_0^\infty dA\, \mathcal{D}(\theta|A\pi)P(A). \qquad (8)
$$

The distribution $P(A)$ is a hyperprior, a prior distribution placed upon a parameter of the Dirichlet prior. Following the same mathematics as Eqs. 5-7, we find that the posterior distribution is the Dirichlet mixture

$$
P(\theta|n) = \int_0^\infty dA\, \mathcal{D}(\theta|A\pi+n)P(A|n)\,, \qquad (9)
$$

where

$$
P(A|n) = \frac{P(A)\mathcal{H}'(n|A\pi+n)}{\int_0^\infty dA\, P(A)\mathcal{H}'(n|A\pi+n)}\,. \qquad (10)
$$

In principle, we have to select and parameterize a functional form for the hyperprior, $P(A)$. For example, an exponential distribution, $P(A) = \lambda\exp(-\lambda A)$, with mean $1/\lambda$, might be appropriate. Fortunately, we can often avoid selecting an explicit hyperprior. In practice, given sufficient data, the probability of that data $P(n|A)$ is a smooth, sharply peaked function of A. This is illustrated in figure 1 using 10^7 observations of the 160,000 dimensional doublet substitution probability, where the mean prior distribution is taken to be the product of singlet substitutions probabilities. If the prior distribution of A is reasonable, and neither very large nor very small over the range of interest, then the posterior distribution $P(A|n)$ will also be very strongly peaked. Moreover, the location of that peak will be almost totally independent of the prior placed on A. In this limit the posterior Dirichlet mixture (Eq. 9) reduces to the single component that maximizes the probability of the data;

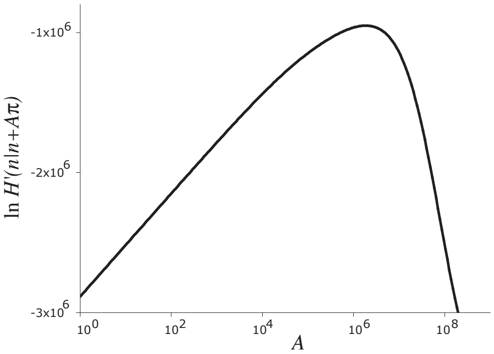

Figure 1: The likelihood of observations as a function of the scale parameter A. With multinomial sampling and a Dirichlet prior the likelihood of the data follows the negative hypergeometric distribution, $H'(n|A\pi+n)$, where n is the count vector of observations, π is the mean prior estimate of the sampling distribution, and A is a scale parameter (Eq. 7). Given a large number of observations (here, $N = \sum n_i$ is about 10^7) the probability of the data is a smooth and very sharply peaked function of the scale parameter A.

$$
\begin{aligned}
P(\theta|n) &\approx \mathcal{D}(\theta|A\pi+n), \\
A &= \mathrm{argmax}_A P(A|n) \approx \mathrm{argmax}_A P(n|A), \\
P(n|A) &= \mathcal{H}'(n|A\pi+n). \qquad (11)
\end{aligned}
$$

Here, $\mathrm{argmax}_x f(x)$ is the value of x that maximizes that function $f(x)$.

Given any function of θ, the average of the function across the posterior distribution (the posterior mean estimate (PME) or Bayes' Estimate) minimizes the mean squared error of that estimate. In particular, the posterior mean estimate of θ (Eq. 4) is

$$
\theta_i^{\mathrm{PME}} = \frac{A\pi_i + n_i}{A+N}. \qquad (12)
$$

Taken altogether, our practice is to take the raw doublet substitution counts and construct a mean prior distribution π based upon the approximation that substitutions on neighboring sites are uncorrelated. We then find the scaling factor A that maximizes the negative hypergeometric probability $\mathcal{H}'(n|A\pi + n)$. For our data the total number of observations N is around 10^7, for which the optimal scale factor A was found to be about 10^6. The posterior mean estimate of the doublet substitution distribution is then used to construct the doublet substitution matrix. Code for constructing doublet substitution matrices using this procedure and for finding the optimal prior and posterior, given any set of observations and π, a best guess for the true distribution θ, is available from our web site (`http://compbio.berkeley.edu`), along with other code and data for this work. Our programs make extensive use of the Open Sourced GNU Scientific Library (GSL) (Gough, 2003; Matsumoto & Nishimura, 1998).

References

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Biological sequence analysis*. Cambridge University Press.

Gough, B., ed. (2003) *GNU Scientific Library Reference Manual*. 2nd edition,, Network Theory Ltd.

Johnson, N. L. & Kotz, S. (1969) *Discrete Distributions*. John Wiley, New York.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science,* **262** (5131), 208–214.

Matsumoto, M. & Nishimura, T. (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.,* **8** (1), 3–30.

Nemenman, I., Shafee, F. & Bialek, W. (2001). Entropy and inference, revisited. arXiv:physics/0108025.